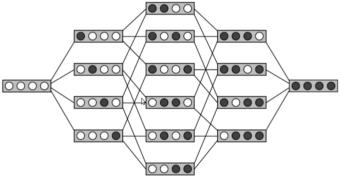
Supervised Learning

Embedded Feature Selection



Learning goals

- Add Learning goals

EMBEDDED FEATURE SELECTION

- Embedded techniques are methods that integrate feature selection directly into the learning process.
- They use an internal criterion of the applied learner to have better control over the search for useful features.
- Embedded techniques usually need to be tailored to each learner.

SVM: RECURSIVE FEATURE ELIMINATION (RFE)

- RFE is a popular backward search technique for the linear SVM and the 2-class problem.
- Here we will always assume standardized features. (This should be the case for an SVM anyway!)
- Coefficient size $|\theta_j|$ tells us how much impact feature j has on our classification, since $f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$ is the decision function.

Idea: Sequentially drop the feature with the smallest $|\theta_j|$.

SVM: RECURSIVE FEATURE ELIMINATION (RFE)

Recursive feature elimination

- Standardize the data.
- Start with full set of features S.
- Fit a linear SVM, using the features in S, and estimate coefficients $\boldsymbol{\theta}$.
- Remove feature(s) j with minimal $|\theta_j|$ from S.
- Iterate using the reduced S for the SVM.

SVM: RECURSIVE FEATURE ELIMINATION (RFE)

Some notes:

- Strictly speaking, this procedure does not perform selection, but rather constructs a ranking of the features.
- As for filters, we need an extra criterion for termination / selection.
- To improve speed one can drop k features in each iteration.
- Extensions to other kernels or multi-class tasks are not trivial.

L1 PENALIZATION / LASSO

LASSO: least absolute shrinkage and selection operator

- As introduced before, linear methods that regularize the coefficients of the model with an $L1$ penalty in the empirical risk

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}\right)^2 + \lambda \sum_{j=1}^{p} |\theta_j|$$

  are very popular for high-dimensional data.
- The penalty summand shrinks the coefficients towards 0 in the final model.
- Many (improved) variants: group LASSO, adaptive LASSO, ElasticNet, ...
- Has some very nice optimality results: e.g., compressed sensing.